From Persistent Identifiers to Digital Objects to Make Data Science More Efficient

Peter Wittenburg†

Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

Keywords: Big data; Data management; Persistent identifiers; Digital objects; Data infrastructure; Data intensive science

Citation: P. Wittenburg. From persistent identifiers to digital objects to make data science more efficient. Data Intelligence 1(2019), 6-20. doi: 10.1162/dint_a_00004
Received: May 7, 2018; Revised: May 30, 2018; Accepted: June 4, 2018

ABSTRACT

Data-intensive science is reality in large scientific organizations such as the Max Planck Society, but due to the inefficiency of our data practices when it comes to integrating data from different sources, many projects cannot be carried out and many researchers are excluded. Since about 80% of the time in data-intensive projects is wasted according to surveys we need to conclude that we are not fit for the challenges that will come with the billions of smart devices producing continuous streams of data—our methods do not scale. Therefore experts worldwide are looking for strategies and methods that have a potential for the future. The first steps have been made since there is now a wide agreement from the Research Data Alliance to the FAIR principles that data should be associated with persistent identifiers (PIDs) and metadata (MD). In fact after 20 years of experience we can claim that there are trustworthy PID systems already in broad use. It is argued, however, that assigning PIDs is just the first step. If we agree to assign PIDs and also use the PID to store important relationships such as pointing to locations where the bit sequences or different metadata can be accessed, we are close to defining Digital Objects (DOs) which could indeed indicate a solution to solve some of the basic problems in data management and processing. In addition to standardizing the way we assign PIDs, metadata and other state information we could also define a Digital Object Access Protocol as a universal exchange protocol for DOs stored in repositories using different data models and data organizations. We could also associate a type with each DO and a set of operations allowed working on its content which would facilitate the way to automatic processing which has been identified as the major step for scalability in data science and data industry. A globally connected group of experts is now working on establishing testbeds for a DO-based data infrastructure.

† Corresponding author: Peter Wittenburg (Email: Peter.Wittenburg@mpi.nl; ORCID: 0000-0003-3538-0106).

1. DATA INTENSIVE SCIENCE IS REALITY

In large research organizations such as the Max Planck Society (MPS) *data-driven science* is being practised for many years. During the last years, however, data-driven science mutated to *data-intensive science* as it was introduced by J. Gray as the 4th Paradigm [1]. Many examples could be given which include all kinds of disciplines. Here we restrict ourselves to three from material science, humanities, and neuro science which indicate the big changes that currently take place.

The NoMaD project, funded by the European Commission, in material science collects billions of results of simulations from different theoretical labs worldwide with the intention to find new types of categorizations. M. Scheffler, Director at the Fritz Haber Institute of the MPS, recently stated: "*Big data, machine learning, and artificial intelligence are revolutionizing numerous fields, and materials science is no exception. Aggregating and analyzing the billions of data from computational materials science will enable us to categorize the endless space of materials and exploit hidden relationships for basic research and societal important applications*." All aggregated material is openly available for analytical purposes.

The DOBES project on documenting endangered languages, run at the Max Plank Institute for Psycholinguistics and funded by the Volkswagen Foundation, engaged 75 international teams of linguists, biologists, ethnologists, and musicologists to record and describe about 100 languages worldwide. All material has been aggregated into one repository making the data available for research purposes. Various repositories with large collections are collaborating representing an enormous resource enabling new kinds of research questions such as "can we better understand the evolution of languages over thousands of years", and "what kind of features in languages compensate for flat intonation contours or small set of sounds?".

In the domain of neurosciences large initiatives such as the human brain project (Figure 1) are working on methods that allow drawing relations between phenomena of brain diseases with patterns that can be found in different types of data sets such as from gene sequencing, brain imaging, etc. In times where an increasing number of people are suffering from brain diseases a deeper understanding about their causes and early detection possibilities are urgently needed.

With respect to these and other examples we can make a number of observations:

- Data are not only created anymore to write scientific papers, but they are created with the notion of being reused in different contexts which is revolutionary in many disciplines;
- Advanced statistical methods such as machine learning are required to detect the patterns and correlations hidden in the data—a modern version of a microscope;
- Data from different labs are needed to fit all free parameters of the underlying models;
- An efficient data infrastructure is required to aggregate, manage, and process all data;
- The efforts for data-intensive science are huge and only advanced labs with sufficient resources can currently carry out such work.

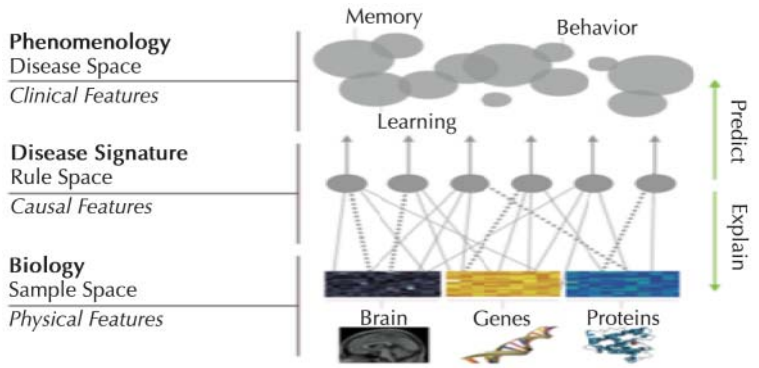

Figure 1. The human brain project[①].

2. DATA REALITY AND TRENDS

A number of studies indicate that data science is suffering from huge inefficiencies. A survey carried out within the Research Data Alliance (RDA) Europe project in 2013 [2] indicated that 75% of the time of data scientists is wasted on data wrangling which includes all those steps that need to be taken before the real analytic work can start (negotiating reuse, finding data and metadata, aggregating, improving quality, transforming formats, etc.). In 2014 M. Brodie reported about a study from Massachusetts Institute of Technology (MIT) [3] according to which 80% of the time of data scientists is wasted on data wrangling and recently (2017) a survey in industry from CrowdFlower [4] reported about 79% of wasted time. In addition, X. Clever reported at the Big Data and AI Summit in 2018[②] that 60% of industrial data projects fail.

The biggest factors for these inefficiencies seem to be a bad quality of data/metadata and a bad data organization, i.e., if a rough description of data is available, it is difficult to find out how to access them, where the corresponding metadata is to enable processing, etc. On top of these inefficiencies there are of course the semantic challenges which can emerge when agreed metadata sets need to be defined, mapping of parameter spaces needs to be done, and knowledge has been extracted and transformed for example into formal assertions which then need to be integrated. Data scientists typically mention the following priorities to be addressed:

- A higher degree of interoperability is required to overcome the huge fragmentation;
- Data scientists are confronted with too much detail in an increasingly complex data and tool landscape;
- Data scientists need wide scale data tracing and reproducibility mechanisms to facilitate trust and verification;
- Improved ways are needed to automatically create scientific annotations and assertions to capture and exploit knowledge.

① https://www.humanbrainproject.eu/en/
② https://www.bitkom-bigdata.de/de/speaker/clever

The result of all these inefficiencies is that much data-intensive science work cannot be carried out and that many researchers are excluded. Obviously the same holds for industry.

According to Intel we will see more than 50 billion smart devices being deployed in the coming years covering all aspects of life. All these devices whether being built into cars measuring driving parameters or being mounted as wearables measuring the state of health of individuals will create huge amounts of continuous high resolution data streams. These data will be used in science as well as in industry and it is obvious that there will be much re-purposing of such data in various contexts.

Given these trends we can conclude that we are not fit to meet these challenges, since the current data practices are not scalable. Urgent measures need to be taken to define standards and an eco-system of infrastructures guaranteeing much more efficiency.

3. PERSISTENT IDENTIFIERS AND METADATA AS BASIC STEPS

In 2008 the first version of the Data Seal of Approval[3] was presented as a light-weight certification scheme for repositories and the data they are hosting to increase trustworthiness. It contains 16 guidelines covering aspects such as findability, accessibility, use of known formats, reliability of data, and association of data with persistent identifiers (PIDs). In 2009 International Council for Science (ICSU) created the World Data Systems Group[4] which as one of its tasks also created a set of criteria to certify repositories and thus increase the level of trustworthiness. In July 2017, the Data Seal of Approval (DSA) and the World Data System (WDS) joined forces under the umbrella of RDA and created the CoreTrustSeal certification criteria[5]. Although CoreTrustSeal is now a global standard, it still needs time to be accepted globally. For completeness, we should mention that in 2012 the NESTOR group formulated the DIN 31644 Criteria for trustworthy digital archives and that in 2012 the ISO 16363:2012 standard[6] was published.

At the DAITF side workshop of the ICRI conference[7] in March 2012 in Copenhagen which was the first meeting toward forming the Research Data Alliance, L. Lannom presented four layers of working with data called Discoverable, Accessible, Interpretable, and Reusable (DAIR). These classifications were the basis for many of the groups in RDA[8]. In May 2013 the Data Working Group of the G8+O6 Group of Senior Officials on Research Infrastructures wrote a white paper with five Principles for an Open Data Infrastructure [5] which were partly inspired by L. Lannom's layers, but including as well the management/curation dimension and the relevance of experts: Discoverable, Accessible, Understandable, Manageable and People.

[3] http://datasealofapproval.org/en/
[4] https://www.icsu-wds.org/
[5] https://www.coretrustseal.org/
[6] https://www.iso.org/standard/56510.html
[7] http://www.icri2012.dk/www.ereg.me/ehome/index06e1.html
[8] https://www.rd-alliance.org/

These discussions lead to the well-packaged FAIR principles [6] formulated first in 2014 and published in 2016 which are now accepted globally and which should be seen as guiding our data practices. These principles summarize the importance of PIDs and MD excellently in their first two dimensions:

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

To be Accessible:

A1. (meta)data are retrievable by their identifiers using a standardized communications protocol.

In 2014 the RDA Data Foundation & Terminology (DFT) Group finalized their work on a core data model[9] and the conceptualization on basic terminology. It went a step ahead of the principles by also writing a model of how these different entities need to be related. In the core is the notion of a Digital Object (DO) which is represented by a structured bit sequence (content) being stored in some repositories, is referenced by a PID and described by metadata (Figure 2). A DO can be a simple data entity or a complex collection. Important is that metadata (MD) and collections are also DOs, i.e., they are assigned a PID and where applicable associated with MD.

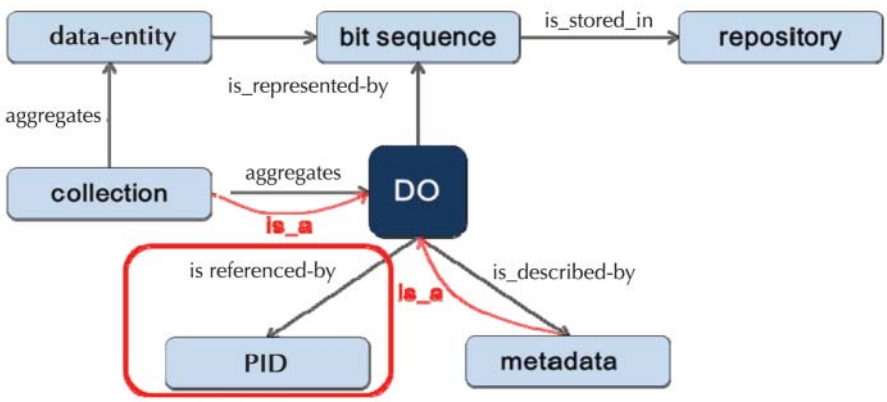

Figure 2. The notion of a Digital Object (DO). Note: PID refers to persistent identifier.

The DFT Group also formulated a suggestion of how these different entities could be linked together to make data processing efficient and reliable. This binding concept is indicated in Figure 3. When we assume that the PID and the PID resolution system are persistent then it makes indeed sense to include crucial binding information in the PID record. Crucial in this context are not just the references to the locations

[9] http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

where the bit sequences are stored and where the metadata can be found, but also for example what the checksum is to be able to quickly prove identity, where the rights record can be found to efficiently do authorization for distributed locations and where the blockchain entry is located that includes transactions and usage agreements. Different repositories formulated wishes for such "kernel" attributes and another RDA group is busy standardizing them[⑧] which would enable global interoperability for this crucial part of data organizations and thus making the specifications machine actionable.

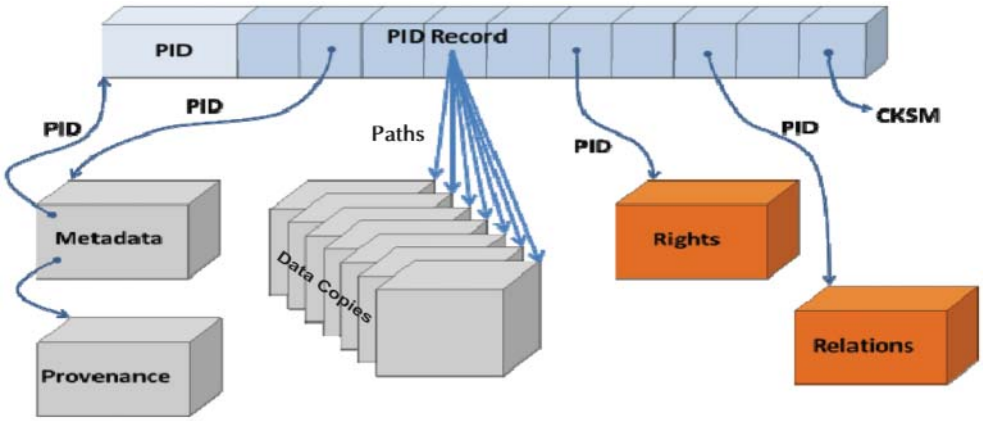

Figure 3. The binding concept of persistent identifier (PID) records.

Recently the Group of European Data Experts (GEDE) collaboration[⑨] including delegates of 47 large European research infrastructures finalized their report on PIDs and their usage[⑩]. After a year of intensive discussions the experts agreed not only on the usage of PIDs, in nearly all infrastructures Handles and DOIs are used—the latter also being Handles with prefix 10. Also agreements were made about the granularity with which PIDs should be assigned, how versioning could be dealt with by PIDs, and other topics. Basically the group of experts also agreed that the PID should be associated indeed with "kernel" attributes including crucial state information.

In this paper we will not make many statements about metadata except stating that there is now much agreement that metadata should be sufficiently rich, that the schema and the used concepts need to be registered in open registries using well-known formal languages (XML, RDF, SKOS, etc.) and that the metadata records should be offered for harvesting via standardized protocols (OAI PMH, ResourceSync, etc.). This explicitness will not solve the interoperability problems, but it will allow interested experts to interpret and reuse the data and carry out semantic mappings if needed. In addition, being able to export metadata as RDF triples opens the way toward Open Linked Data and related work[⑪].

⑧ https://www.rd-alliance.org/groups/pid-kernel-information-wg
⑨ https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
⑩ https://zenodo.org/record/1116189
⑪ http://linkeddata.org/

4. DEPENDENCE ON PIDS

Using PIDs offers thus a number of great advantages such as clear and stable identities allowing humans and machines to exactly refer to the right data even after many years, to have easy ways to prove identity, integrity, and authenticity, to provide stable references also as basis for citations, to easily find descriptive metadata, and information needed for authorization, for reuse tracing information, on versioning, etc. We realize, however, that we are increasingly dependent on a stable PID system that includes a stable, robust, and globally available resolution system that should be independent of any protocol since these will change over time. As Figure 4 suggests PIDs can become as crucial for the data domain as IP became for the Internet. Some argue that we could have a variety of PID systems for registering Digital Objects risking the typical fragmentation problems. Others argue that as in the case of TCP/IP, having a harmonized numbering system and a unified protocol would be best as a starting point for a global domain of registered DOs and a new phase of rich exploitation opportunities. I believe that we have the basis for such a global PID system that will need technological improvements over the years to meet the challenges.

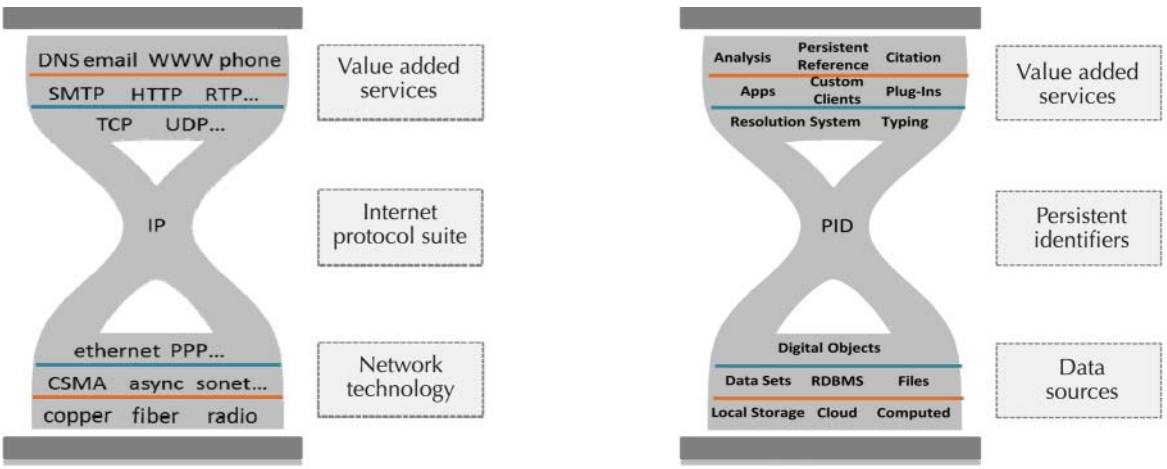

Figure 4. Persistent identifiers (PIDs) crucial for the data domain.

From the various PID systems that have been invented during the last 20 years only the Handle System® found the broad support which is needed to provide stable services. A Handle has a prefix which is assigned by an authorized registration authority, while the suffix can be specified locally (Figure 5). The string is based on Unicode 2.0 with UTF-8 encoding and for the suffix there are no substantive restrictions which also mean that the address space is gigantic. DOIs® have the prefix "10.x". It is widely recommended not to include semantics into the string, since it may have a meaning for the production process, but will not be meaningful for the consumption of data in different contexts.

® https://en.wikipedia.org/wiki/Handle_System
® https://en.wikipedia.org/wiki/Digital_object_identifier

Figure 5. An example of a Handle.

The Handle System offers a robust and highly performant resolution system based on stable specifications. The Swiss DONA Foundation® with its international board is in charge of maintaining the system and guiding it through the coming decades. The DONA board is also in charge of guiding the Global Handle Registry (GHR) which is now based on a distributed and redundant network of so-called Multiple Primary Administrators (MPAs) which run a root resolver, adhere to special rules defined by DONA and also act as registration authorities. Currently, the GHR has 10 globally distributed MPAs with a broad continental coverage—one residing in China. Figure 6 indicates the set-up of the Handle system with the distributed GHR in the core. A calculation indicates that the DONA Foundation will be able to carry out its work self-sustained when the number of MPAs which pay an annual fee has reached 12, i.e., financial stability of this service will be reached soon.

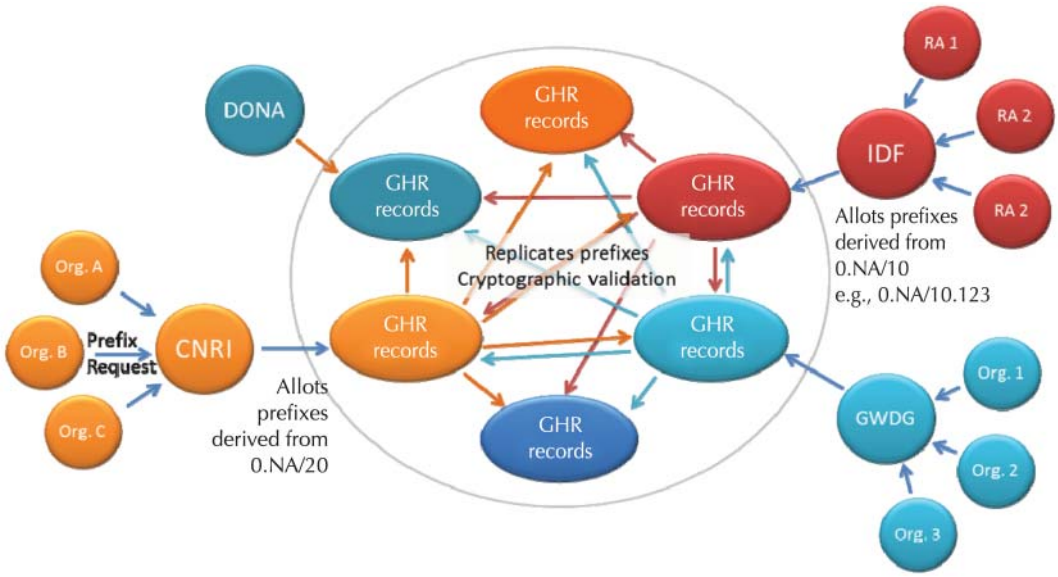

Figure 6. Global Handle registry system.

® https://www.dona.net/

5. DIGITAL OBJECTS

We define Digital Objects as "meaningful entities" within a specific application domain existing in the digital world of bits. Their content can include data, collections, metadata, software, publications, configurations, categories, assertions, or other digital entities. "Meaningful" here means that people want to talk about data, work with them, process them, refer to them, or cite them. In addition, we follow the specifications of the Data Foundation & Terminology Group mentioned above: it is represented by a structured bit sequence being stored in some repositories, is referenced by a PID and described by metadata. The PID record is used to bind crucial entities together.

Already in 1995 Kahn and Wilensky [7] introduced the term "Digital Object" (DO) and a framework for distributed services for DO (Figure 7). According to them a DO has a structured bit sequence, a persistent ID and key metadata with at least one key-value pair covering the PID. In 2006 they renewed their paper and referred to a Digital Object Architecture being developed at CNRI [8]. They discovered that after the introduction of the Internet something essential was missing. While the Internet specifies how messages, in general not having a meaning, are being exchanged between network devices. Finally, however, the purpose is to exchange "meaningful entities" between repositories or processing units. This was the motivation to define the concept of "Digital Objects" as meaningful entities being subject of exchange and processing. Digital Objects are thus central for human and machine communication and the need to identify and describe them to enable interpretations.

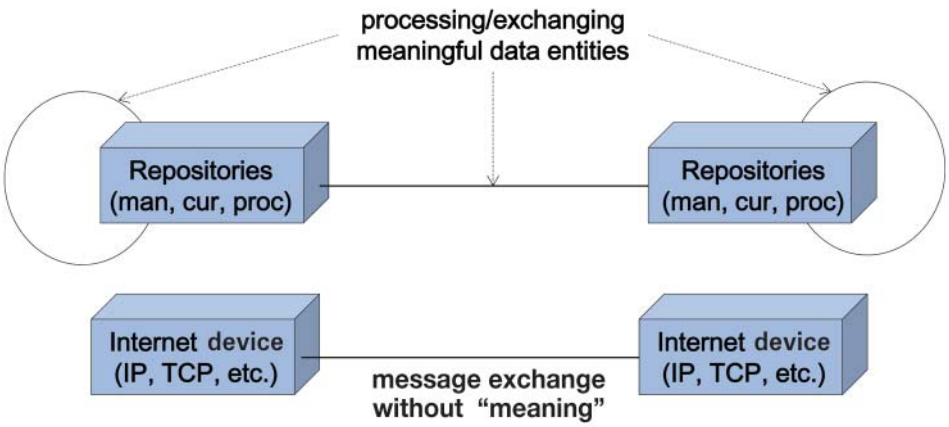

Figure 7. The concept of Digital Object.

Consequently, we should define a Digital Object Access Protocol (DOAP)[@] [9] that enables these exchange and processing steps in a way that is independent on how repositories organize their data (Figure 8). Repositories could store their data in files, clouds, SQL, and no-SQL databases and the way they relate data and different types of metadata with each other can again be very different. A unified DOAP

[@] Some experts use the term Digital Object Interface Protocol which in this paper is seen as a synonym.

would help to bridge across all these differences and achieve interoperability at the level of data organizations reducing the costs considerably. Some repositories would have to develop adapters to meet the DOAP requirements, and others may have chosen for an internal organization perfectly matching the DOAP requirements. It is urgent time to design this DOAP as a global effort as it is planned in the C2CAMP collaboration®.

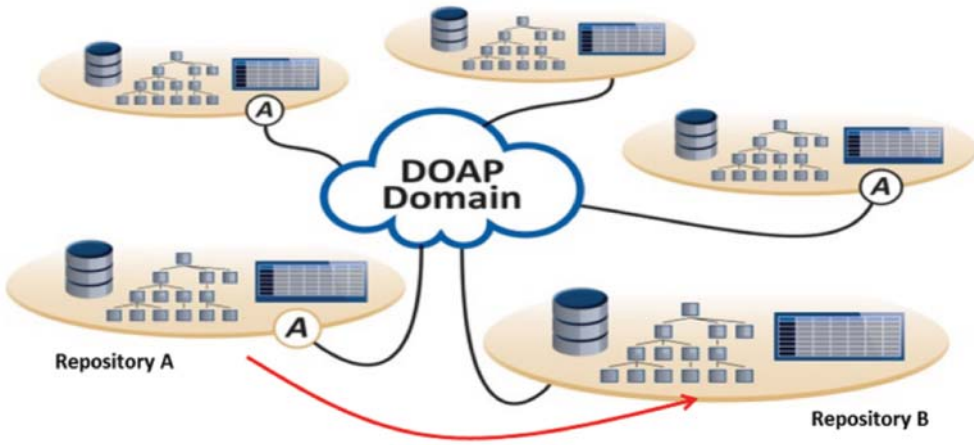

Figure 8. The concept of Digital Object access protocol.

But we can make another step ahead by looking into the way "objects" have been introduced in computer science and practical information technology. In 1974 the concept of "abstract data types" was introduced to support encapsulation, to localize change and to achieve flexibility with respect to changing implementations. A user point of view is taken which defines the behavior of a specific type of object simply by a set of values and a set of operations on them. This concept was extended to the concept of "object-oriented-programming" where "methods" are defined to manipulate the state of the object and where objects are interacting with one another. It is known that these concepts facilitated the path to designing and implementing complex systems.

Later the notion of "objects" was used to describe stores where the user does not deal anymore with physical structures such as directory paths and files, but where a virtualization layer is being offered defined by PIDs and some metadata. Cloud systems are implementing such a virtualization and hash values linked with the metadata enable the access to the bit sequences independent of where and how they are stored. Operations such as "upload" act on the virtualization layer and are translated internally into well-tested procedures that work on concrete structures. However, the administration layers of these solutions including the PIDs (Hash values) being used are only locally valid.

® http://github.com/c2camp/core/wiki

With a globally resolvable PID system and a binding concept in place we can now do this next step borrowing the idea of "objects" which goes back to the definition of abstract data types. Complex metadata defining a specific type of data entities can be summarized into an "Object Type" where each instance is being defined by a globally resolvable PID. By introducing a Data Type Registry[⑨] that allows researchers and others to link types with all kinds of operations we come close to the extension of our Digital Object concept as introduced at the beginning of this section. For all object types the set of valid operators for their manipulation, their visualization, etc., is known and can be retrieved, i.e., we can establish a domain that can be compared to those of Multipurpose Internet Mail Extensions (MIME) types, however, offering the flexibility needed for scientific DO types.

6. DO-BASED APPROACHES

A number of experts from about 20 strong institutions and initiatives have now formed the C2CAMP collaboration to turn the DO approach into a test-bed with runnable code where the RDA platform will be used to push further specification work. Two major aspects are being looked at:

- What are the steps needed to come to a Global Digital Object Cloud (GDOC)[⑩], a domain of registered Digital Objects?
- What are the steps needed to come to a Type-Triggered Automatic Processing domain?

The GDOC domain is best characterized by Figure 9 created by L. Lannom. The user only deals with logical representations of the Digital Objects, i.e., in his or her virtual research environment on his or her notebook that he or she operates in a world of metadata descriptions and PIDs. The corresponding services such as a PID resolution service, search portals for descriptive metadata, registries to include authorization records, etc. are offered by different types of repositories and registries. These services represent a virtualization layer on top of concrete implementations such as repositories which could have been setup using different data organizations such as clouds, file systems, and SQL, or no-SQL databases. The user is thus working on a level which is suited to his or her domain of conceptualization which includes logical descriptions of the Digital Objects and hides organizational complexity.

One concrete task for the C2CAMP collaboration will be to develop typical operations such as "move a DO", "copy DO's bit sequence", "delete DO's bit sequence", etc. which operate in the global domain of Digital Objects all based on an existing DOAP and repositories/registries supporting it.

⑨ https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries
⑩ http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3

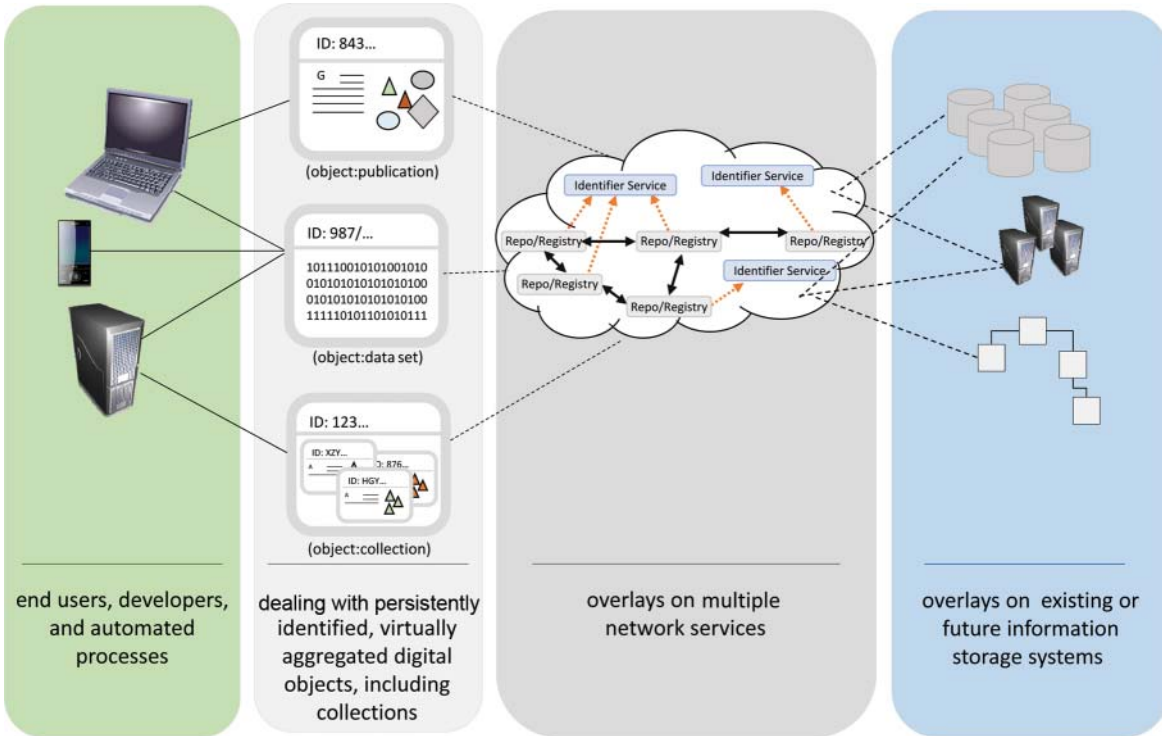

Figure 9. The characteristics of a Global Digital Object Cloud (GDOC) domain.

The second aspect to be tackled is to make optimal use of the Object Type and its known operators specified in the Data Type Registry for the creation of automatic workflows acting in a domain of DOs which is too big to be scanned manually. The concept of Type-Triggered Automatic Processing (T-TAP) as it is indicated in Figure 10 assumes that data creators or stores announce that they have new data by exposing their PIDs and their metadata to a structured data market. Software agents that are fed with profiles describing useful data act on behalf of researchers and trying to match with the metadata of the offered data. In case of a positive match automating processing steps are being executed which lead to new results. The researcher is informed about these results and can trigger visualizations or further analysis steps to take decisions. The basis for implementing such advanced T-TAP scenario is the implementation of the DO concept including crucial entities such as DOAP, Data Type Registries, and a functioning structured market place with sufficiently rich metadata descriptions and types of metadata that support automatic processing.

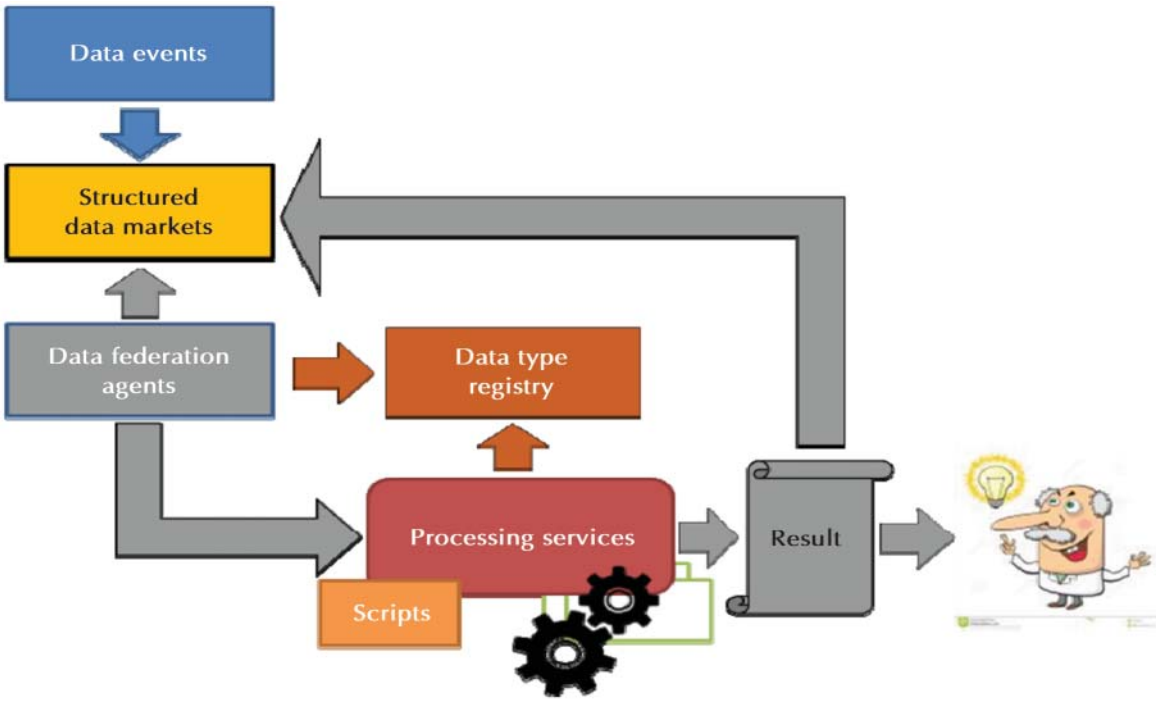

Figure 10. The concept of Type-Triggered Automatic Processing (T-TAP).

The C2CAMP collaboration has been started to act on these two major aspects and is open for further collaboration interests. It includes currently data scientists from a number of research communities (environment, languages, biodiversity, and climate modeling), and some large data and HPC centers closely collaborating with such communities, some IT centers, and service providers. The claim is to contribute to an emerging larger test-bed to showcase the two major aspects and to integrate useful and compliant running code.

7. CONCLUSIONS

After intensive discussions in the realm of the Research Data Alliance and also in the International Telecom Union, which led to a number of PID related specifications and clarifications about the use of PIDs, a group of international data experts started working on a Digital Object based infrastructure making use of the specifications so far, but also drive further specifications and the implementation of components. Digital Objects will enable users to ignore all the details about how repositories exactly store the data and metadata and how they relate the various information entities which are necessary to manage and process data. They can operate at the level of logical representations of their data by using PIDs and metadata which is their conceptual level. In addition, by associating types with DOs and linking types to sets of operations

the way toward increasing the amount of automation is opened. This group (C2CAMP) including experts from various disciplines and backgrounds is committed to developing codes that can be fed into an extendable testbed to demonstrate the power of the concept. The group does not have to start from scratch, since essential elements have already been implemented in first versions. The group is open for other implementers, but also for experts who are interested in discussing the concept, proving its applicability in different contexts as a recent initiative in Europe has shown that it found more than 100 active supporters who want to contribute.

REFERENCES

[1] T. Hey, S. Tansley, & K. Tolle. The fourth paradigm: Data-intensive scientific discovery. Available at: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[2] RDA EU Survey. Available at: http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f.

[3] M.L. Brodie. Understanding data science: An emerging discipline for data-intensive discovery. In: Proceedings of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID'2015), 2015, pp. 238–245. Available at: http://pdfs.semanticscholar.org/46b2/bddce0f3b47f8b9dbe9e05777db9a24d8c39.pdf.

[4] CrowdFlower survey. Available at: https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html.

[5] A. Blatecky, J. Bicarregui, & C.M. Pires (eds.) G8 principles for an open data infrastructure. Available at: http://purl.org/net/epubs/work/12236702.

[6] FAIR Principles. Available at: https://www.force11.org/group/fairgroup/fairprinciples.

[7] R. Kahn, & R. Wilensky. A framework for distributed digital object services (1995 version). Available at: http://www.cnri.reston.va.us/k-w.html.

[8] R. Kahn, & R. Wilensky. A framework for distributed digital object services (2006 version). Available at: https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf.

[9] P. Wittenburg, & G. Strawn. Common patterns in revolutionary infrastructures and data. Available at: https://b2share.eudat.eu/records/4e8ac36c0dd343da81fd9e83e72805a0.

AUTHOR BIOGRAPHY

Peter Wittenburg was Executive Director of Research Data Alliance (RDA) Europe, a member of RDA Technical Advisory Board, and Scientific Coordinator of European Data Infrastructure (EUDAT). He set up and led the Technical Group with about 30 experts at Max Planck Institute (MPI) for Psycholinguistics and then led the Language Archiving Group with about 25 experts. He has set up experimental labs at the MPI for Psycholinguistics including a variety of channels (speech, intonation, gesture, eye tracking, electroencephalogram, etc.), developing a variety of analysis and digital signal processing tools including speech processing and speech perception simulations. From 1988 to 2012 he was a member of the Central IT Advisory Board of the Max Planck Society (MPS) and in 2011/12 also member of the IT Strategy Committee of the MPS. From about 1995 until 2012 he led the set-up of an archive for language resources and the development of the Language Archiving Technology tool set. Since 2000 he has played leading roles in a variety of European (funded by the European Commission (EC)) and national projects (funded by MPS, DFG, BMBF and NWO[2]) and ISO initiatives (ISO TC37/SC4). From 2000 to 2011 he was responsible for the archiving and accessibility of endangered languages documented in the DOBES program funded by the Volkswagen Foundation. In 2010/11 he was a member of EC's High Level Expert Group on Scientific Data (Riding the Wave report). From 2011 to 2014 he was a scientific coordinator of the EUDAT data infrastructure and the DASISH SSH cluster project. From 2011 to 2012 he was a member of the Steering Board of RDA. He won the Heinz Billing Award of the MPS for the advancement of scientific computation in 2011 and received an honorary doctorate from University Tübingen in 2013.

[2] DFG: German Research Foundation, BMBF: Bundesministerium für Bildung und Forschung, and NWO: The Netherlands Organisation for Scientific Research.